Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W483–W488***
doi:10.1093/nar/gki428

SNPAnalyzer: a web-based integrated workbench for single-nucleotide polymorphism analysis

Jinho Yoo[1], Bonghee Seo[1] and Yangseok Kim[1,2,*]

[1]Bioinformatics Unit, ISTECH Inc., No. 506, Woongshin Art Plaza, 847 Janghang2-dong, Ilsan-gu, Goyang-si, Gyeonggi-do, 411-837, Republic of Korea and [2]Cancer Metastasis Research Center, Yonsei University College of Medicine, 134 Shinchon-dong, Seodaemun-gu, Seoul 120-752, Republic of Korea

Received January 19, 2005; Revised March 9, 2005; Accepted March 24, 2005

ABSTRACT

SNPAnalyzer is a software that performs four essential statistical analyses of SNPs in a common computational environment. It is composed of three main modules: (i) data manipulation, (ii) analysis and (iii) visualization. The data manipulation module is responsible for data input and output, and handles genotype, phenotype and genetic distance data. To ensure user convenience, the data format is simple. The analysis module performs statistical calculations and consists of four subcomponents: (i) Hardy–Weinberg equilibrium, (ii) Haplotype Estimation, (iii) linkage disequilibrium (LD) and (iv) quantitative trait locus analysis. The main feature of the analysis module is multiple implementations of different algorithms and indices for haplotype estimation and for LD analysis. This enables users to compare separate results generated by different algorithms, which help to avoid biased results acquired by applying a single statistical algorithm. The performance of all implemented algorithms has been validated using experimentally proven datasets. The visualization module presents most of the analyzed results as figures, rather than as simple text, which aids in the intuitive understanding of complex data. The SNPAnalyzer has been developed using C and C++ and is available at http://www.istech.info/istech/board/login_form.jsp.

INTRODUCTION

The analysis of human genetic variation is a key step toward finding disease susceptibility markers and gaining an understanding of different drug responses among individuals. A single nucleotide polymorphism (SNP) is a single base mutation in DNA, and is the simplest form and the most common source of genetic polymorphism in the human genome. Many researchers and scientists have developed various statistical and computational technologies for the extraction of valuable information from the vast accumulation of deposited SNP data. A haplotype is a particular pattern of sequential SNPs or alleles on a single chromosome. Statistical or computational haplotype estimation algorithms have been developed and have been employed in many computer programs. For instance, Clark (1) devised a rule-based algorithm to estimate haplotype frequencies in a given population. The expectation–maximization method (2) was used to develop a statistically confident algorithm by Excoffier and Slatkin (3). Other elegant algorithms have been developed using a Bayesian approach, which allows both the reconstruction of an individual haplotype and an estimation of haplotype frequencies in a given population (4,5). A considerable amount of research on association has been conducted, where the main focus is to identify the relationships between genotypes and phenotypes. For instance, Nakaya *et al*. (6) studied the relationship between multiple genetic factors with oral glucose tolerance using the OLETF model rat. Many computer programs have been developed to accomplish linkage and association studies by employing the previously stated algorithms. For instance, EH developed by Terwilliger and Ott (ftp://linkage.rockefeller.edu/software/) (7), PHASE developed by Stephens *et al*. (5) and Haplotyper developed by Niu *et al*. (4) are widely used to estimate haplotype frequencies from unphased diploid genotype data. As for linkage disequilibrium (LD) analysis, the genetic linkage distributed within a specific genomic region may be estimated by LD indices. GOLD, the program developed by Abecasis and Cookson (8), is a graphic tool that can display the extents of LD within two different loci as different colors. Quantitative trait locus (QTL) Cartographer (ftp://statgen.ncsu.edu/pub/qtlcart/), used for QTL analysis, can measure the influencing power of loci on a specific trait or phenotype.

While many of the currently available programs are well suited for specific genetic variation studies, such as haplotype

*To whom correspondence should be addressed. Tel: +82 31 9031155; Fax: +82 31 9031152; Email: yskim@istech21.com

estimation, LD and QTL analysis, there are few programs that can perform comprehensive and integrated analyses in a single interface. In addition, existing programs are not entirely convenient for research use owing to several disadvantages. First, many programs still have a text-based, low-quality visual presentation environment. Data conversion is very laborious when the number of SNPs is large. Second, many programs are not very sophisticated in terms of system integration, i.e. several different programs are required to accomplish different genetic variation studies. In order to overcome these disadvantages, we have developed a novel integrated workbench, SNPAnalyzer, which is equipped with multiple algorithms, methods and analysis components for comprehensive analyses of human genetic variations. SNPAnalyzer can also handle a large amount of SNPs in a rather manageable time as compared with currently available programs. The SNPAnalyzer is implemented through web interface with an elegant graphical user interface and it can be freely accessed at http://www.istech.info/istech/board/login_form.jsp after a simple registration process.

COMPONENT DESCRIPTION

The SNPAnalyzer can be accessed through a web interface. All the algorithms employed in SNPAnalyzer were developed with C/C++ and the web interface was implemented using ActiveX.

SNP, genotype and haplotype data

We obtained haplotype data from the dbSNP database at NCBI (http://www.ncbi.nlm.nih.gov/SNP/). Haplotypes are for the HLA gene region where traditional serological alleles have been defined at the molecular (SNP) level. Haplotypes from African American and Asian American ethnic groups were selected for analysis from five possible ethnic groups, being African American, Asian American, Caucasian, Latin American and Native American. The African American group comprises 72 individuals and the Asian American group comprises 75 individuals. Family histories were not considered in our analysis. Each group's haplotypes were redefined by using only biallelic SNPs, and individual's genotypes were reconstructed from these redefined haplotypes. The number of sample data and SNPs are summarized in Table 1. The Hardy–Weinberg equilibrium (HWE) test, the haplotype estimation and the LD test were implemented for these reconstructed genotypes, and accuracies of estimated haplotypes and the patterns of LD were investigated. For QTL analysis, we generated simulation data owing to the lack of quantitative phenotype data.

Table 1. The numbers of individual of each ethnic group and the number of SNPs used for redefining haplotypes

Ethnic group	Afr	Asi
Sample no.	72	75
SNP no.	22	22

African American and Asian American ethnic groups are denoted as Afr and Asi, respectively.

Hardy–Weinberg equilibrium

The HWE component tests whether frequencies of alleles in a specific group keep equilibrium states. Statistics such as chi-square value and *P*-value for the selected loci are displayed in a user-friendly table format. SNPAnalyzer simultaneously executes the HWE test on all loci.

Haplotype estimation

In order to reconstruct haplotypes, we have used several famous algorithms such as the Clarks' algorithm (1), the EM-based algorithm (3) and the Pseudo Gibbs Sampler algorithm (PGS) (5). These algorithms can estimate haplotype frequencies or can reconstruct individual haplotypes from an unphased diploid genotype using statistical or computational methods. The Clark's algorithm (1) is appropriate for the case where samples' genotypes are generally homozygous. The EM-based algorithm (3) uses a likelihood-based method to estimate haplotype frequencies within the given population and consists of two steps: (i) expectation formulation step and (ii) maximization step. The EM-based algorithm can analyze a relatively small number of SNPs because it computes all the possible frequencies of haplotypes within the population. The PGS algorithm (5) uses conditional probabilities and iterative sampling to reconstruct individual haplotypes, and it computes the confidence level of the reconstructed individual haplotypes. These three algorithms are incorporated in the SNPAnalyzer for easier comparison between respective estimation results computed by different algorithms. In addition, unlike other programs, SNPAnalyzer requires minimum data conversion.

Accuracy measure

We employed two methods to measure the accuracy of the haplotype estimation (5): (i) reconstruction of the haplotypes of sampled individuals, which was Clark's (1) main focus and (ii) estimation of sample haplotype frequencies, which was Excoffier and Slatkin's (3) main focus. The first method measures the accuracy with the average error rate, which is the ratio of the number of incorrectly reconstructed samples to the total number of samples. The second method represents the discrepancy between the true haplotype frequencies and the estimated haplotype frequencies by index D, given as $D = \sum_j |\hat{f}_j - f_j| /2$, where $\hat{f}_j$ is the estimated haplotype frequency and f_j is the true haplotype frequency of the j-th sample (5).

Linkage disequilibrium

The extent of genetic association or linkage between two different loci located in a specific chromosome can be estimated by using the recombination fraction (9). There are several indices for detecting LD (10–13). The SNPAnalyzer adopts five indices (D, D', $|D'|$, Δ, Δ^2) and Fisher's exact *P*-value to estimate the extent of LD between two loci. The SNPAnalyzer displays two different graphical outputs for a more sophisticated interpretation of the LD, one of which is a LD map and the other is a four gamete test. It employs an EM-based algorithm to estimate haplotype frequencies.

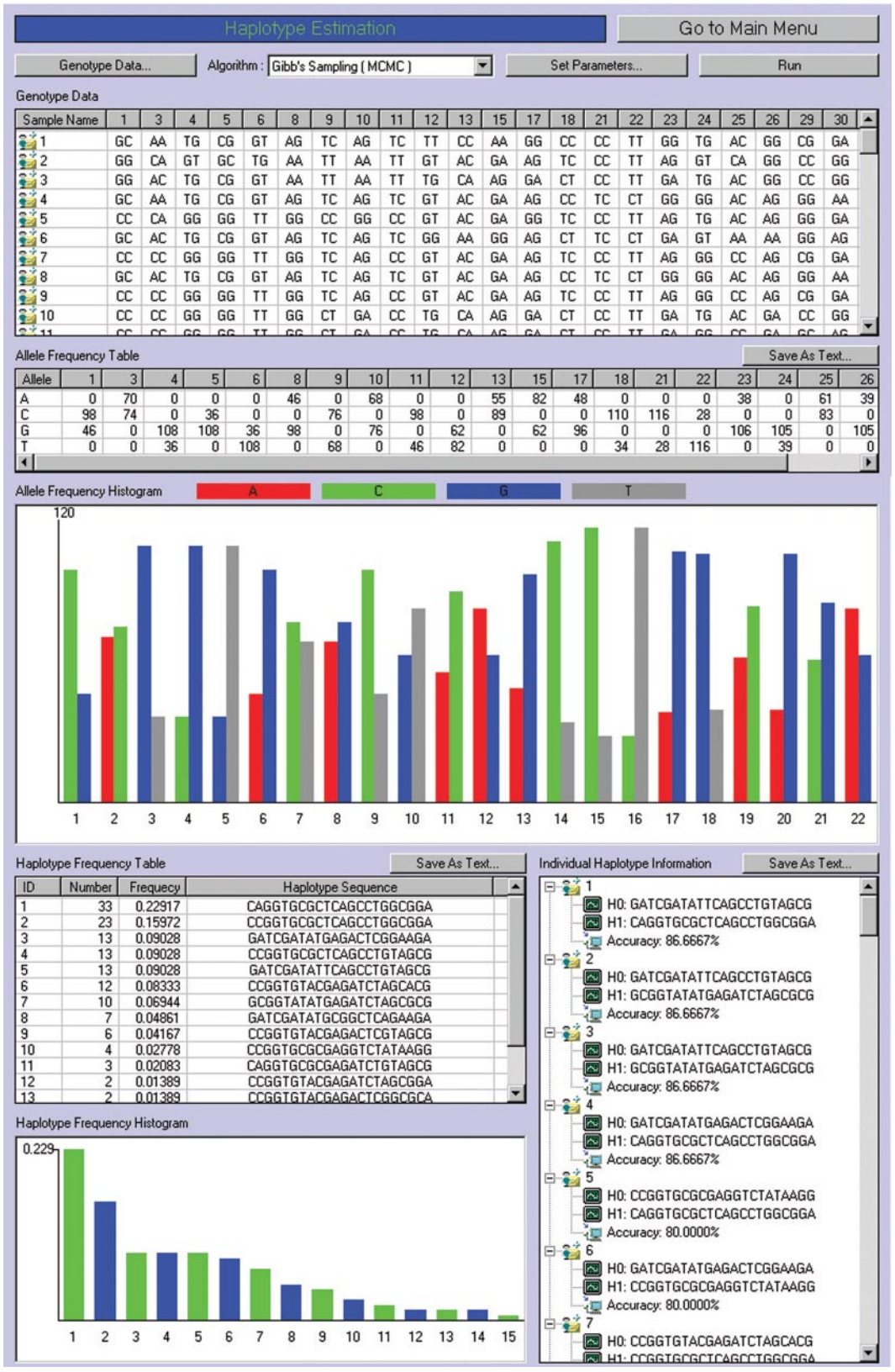

Figure 1. A screen shot of the Haplotype Estimation component. Algorithm selection and data importing are managed at the top panel, which is followed by the contents of the imported individuals' genotypes and the frequencies of the observed alleles. The middle histogram shows the distribution of the frequencies of the observed alleles. Among the three bottom panels, the left panels show the haplotypes' frequencies estimated from genotypes and frequency histograms. The right panel displays the individuals' reconstructed haplotypes and their reconstruction accuracies.

QTL analysis

QTL is a locus which is closely related to a specific trait or phenotype. In SNPAnalyzer, single locus and two loci analyses have been implemented. For single locus analysis, logarithm of odds (LOD) score, which is used globally, is adopted to measure the impact of a locus on a specific phenotype trait. For multi-loci analysis, ANOVA is used to estimate the combined effect of two loci on a specific trait or phenotype (6), where we divided the population group into a maximum number of nine subgroups according to the observed genotypes. The significance of the effect of two major alleles on a specific phenotype is represented by the *F*-ratio (6). The SNPAnalyzer expresses *F*-ratios of all the combined effects of two loci in a two-dimensional map.

IMPLEMENTATION AND EVALUATION

Hardy–Weinberg equilibrium test. The SNPAnalyzer can handle over 300 SNPs simultaneously in the HWE test. Imported genotype data, observed and expected frequencies of genotype, and computed results such as chi-square value and *P*-value are displayed in tables. The SNPAnalyzer uses a significance level of $\alpha = 0.05$ for the analysis.

Haplotype Estimation. The Haplotype Estimation component consists of three parts, which are the control panel, the input data display and the output data display. Data importing and algorithm selection are managed in the control panel, and the observed alleles and genotypes are shown in the input data display by table and histogram after importing data. In the output data display, estimated haplotypes and their frequencies are displayed after completing the analysis, and reconstructed haplotypes of individuals and reconstruction accuracies are also displayed. Figure 1 shows the haplotype estimation results produced by SNPAnalyzer. The SNPAnalyzer can handle up to 24 SNPs simultaneously by using the EM-based algorithm (3), up to 128 SNPs by using the PGS algorithm (5) and more than 200 SNPs by using the Clark's algorithm (1). We applied three different algorithms, Clark's algorithm, the EM-based algorithm and the PGS algorithm, on the African American and Asian American ethnic groups to estimate haplotypes and checked the accuracy of the estimation results. Of the 15 true haplotypes of the African American group, 14 haplotypes were correctly estimated. Only one haplotype was mismatched by both the EM-based algorithm and Gibbs sampling-based algorithm. The frequency differences between the true and the estimated haplotypes were measured by *D*. The *D* scores for the EM-based algorithm and for Gibbs sampling-based algorithm were 0.018 and 0.021, respectively. In the Asian American group, 13 of 14 haplotypes were correctly estimated by the EM-based algorithm and Gibbs sampling-based algorithm separately, and *D* scores for the EM-based algorithm and for Gibbs sampling-based algorithm were 0.033 and 0.017, respectively, which are also negligible. For the reconstruction of haplotypes of individuals, the haplotypes of only two individuals were incorrectly reconstructed by both algorithms for both ethnic groups. However, the analysis result produced by Clark's algorithm was not as accurate as the other algorithms. Table 2 shows the accuracies of the haplotype estimation for the three different algorithms that are employed by SNPAnalyzer.

Table 2. The accuracies of haplotype estimation produced by SNPAnalyzer

Ethnic group	Afr		Asi	
Error type	DIS	AER	DIS	AER
Gibbs	0.021	0.028	0.017	0.027
EM	0.018	0.028	0.033	0.027
Clark	0.156	0.222	0.162	0.293

African American and Asian American ethnic groups are denoted as Afr and Asi, respectively. Gibbs, EM and Clark represent Gibbs sampling-based algorithm, the EM-based algorithm and Clark's algorithm, respectively.

Linkage disequilibrium. The LD component consists of four parts, which are the control panel, the observed data display, the output display and the pattern map. Data importing and loci selection are managed in the control panel. Observed genotypes are displayed in the observed data display. Chi-square statistics, *P*-value and LD indices, such as D, D', $|D'|$, Δ and Δ^2, are computed and displayed in the output display after the LD analysis is completed. These results are followed by an LD pattern map and four gamete test results. Before measuring LD between two loci, haplotypes were estimated by the EM-based algorithm. Figure 2 shows analysis results, one of which is a map of $|D'|$ and the other is a four gamete test result of the African American and Asian American ethnic groups. In the African American ethnic group, $|D'|$ was 1.0 over the genomic region ranging from the 3rd to the 9th SNP and the genomic region ranging from the 10th to the 14th SNP, which shows that there are two separate LD blocks in the specified genomic region. In the Asian American group, there are also two separate LD blocks in the specified genomic region, where one block ranges from the 4th to the 10th SNP and the other ranges from the 11th to the 14th SNP. The LD blocks are seen in the four gamete test results, the patterns of which are similar to those of $|D'|$.

QTL analysis. Single locus and multiloci analyses were implemented using simulation data, which consist of individual genotype data, phenotype data and genetic distance data. The QTL analysis component consists of four parts, which are the control panel, the input data display, the output data display and a graphical view. Users can import genotype, phenotype and genetic distance data at the control panel, and imported data are displayed at the input data display. All estimated results are displayed in the output display after the analysis is complete. Finally, the impact of a single locus or the combined effect of two loci will be shown graphically.

DISCUSSION AND FUTURE PLAN

Most software packages that are currently available provide a single function and are difficult to install and manipulate. Therefore, we have developed SNPAnalyzer to make SNP research more convenient. The main features of SNPAnalyzer can be summarized into four categories: (i) user-friendly interface, (ii) integration of four different analysis processes into a common environment, (iii) multiple implementations of validated algorithms and indices for haplotype and LD analysis and (iv) visualized output. Furthermore, multiple implementations of different algorithms enable users to

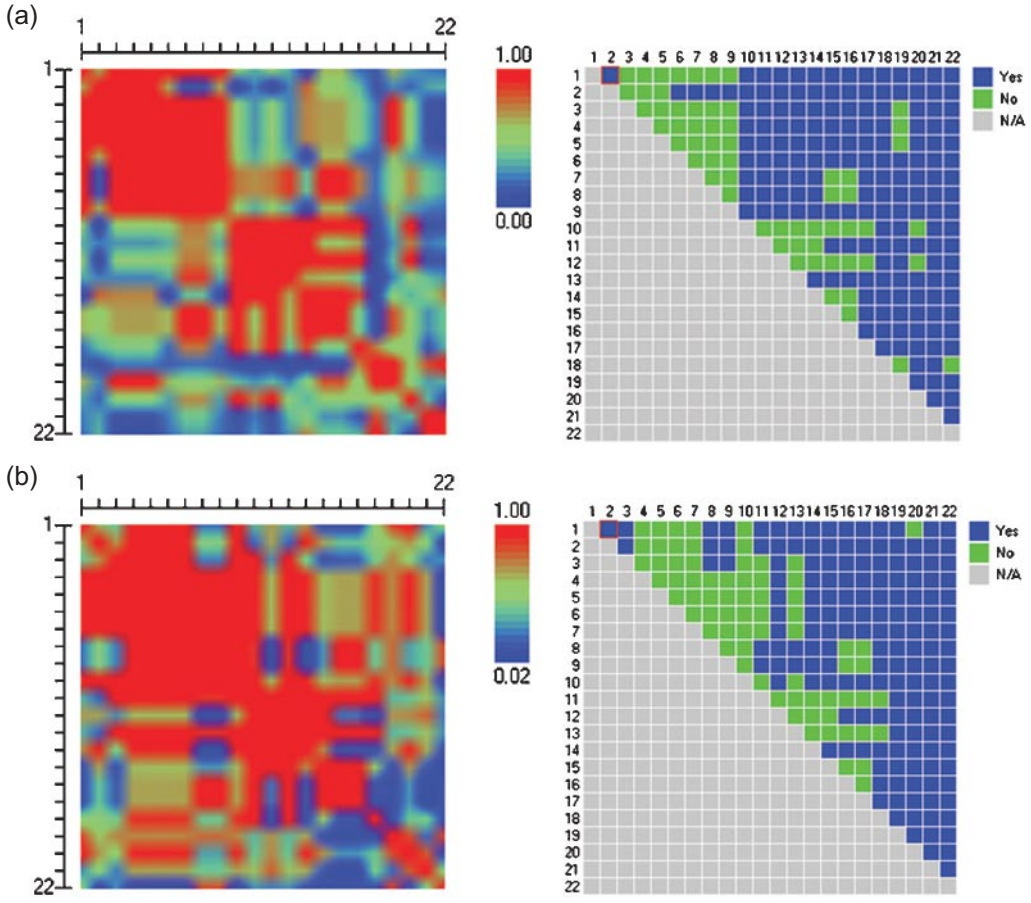

Figure 2. The result of the LD and four gamete tests on two ethnic groups. (**a**) the LD pattern of $|D'|$ of the African American ethnic group reveals that there exists two small LD blocks in the specified genomic region, which are shown in red, and where the value of $|D'|$ is close to 1.0. (**b**) LD pattern of $|D'|$ of the Asian American ethnic group reveals that there exists one small LD block and one relatively large LD block in the genomic region. The right-hand side of the LD pattern displays the results of the four gamete test, the patterns of which are similar to the LD patterns of $|D'|$.

compare results from different analysis methods and to select proper algorithms for their own data. Considering that current SNP studies deal with large amounts of data, good visualization is essential to knowledge discovery when using complex datasets.

The implemented algorithms were validated by using data of 72 individuals of African American ethnicity and 75 individuals of Asian American ethnicity. Data were gathered from the dbSNP database at the NCBI. The analyzed results were evaluated by comparing them with experimentally derived haplotype data according to two accuracy measure categories (5). The EM-based algorithm and Gibbs sampling-based algorithm estimated and reconstructed haplotypes of both ethnic groups with high accuracy. LD patterns within the specific genomic regions were roughly similar in both ethnic groups, both of which showed two LD blocks.

A future version of SNPAnalyzer will be capable of analyzing multiallelic SNP data or microsatellite data, and will be complemented by some additional modules that will allow for more sophisticated analyses. Haplotype blocks refer to sites of closely located SNPs which are inherited in blocks, and haplotype tagging SNPs are a minimal subset of SNPs that can characterize the most common haplotypes. Because the current genetic association study identified an increased importance of haplotype blocks and haplotype tagging SNPs, the next version of SNPAnalyzer will incorporate a haplotype blocking module by using at least two state-of-the-art algorithms. In addition, several statistical methods, such as cross-tabulation analysis and logistic regression analysis, will also be developed to facilitate case–control or population-based association studies.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by grant from Korea Science and Engineering Fund (KOSEF) through the Cancer Metastasis Research Center at Yonsei University College of Medicine (Project Number M1-0218-00-0016). Funding to pay the Open Access publication charges for this article was provided by ISTECH Inc., Republic of Korea.

Conflict of interest statement. None declared.

REFERENCES

1. Clark,A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.*, **7**, 111–122.
2. Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.
3. Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.
4. Niu,T., Qin,Z.S., Xu,X. and Liu,J.S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **70**, 157–169.
5. Stephens,M., Smith,N.J. and Donnelly,P. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, **68**, 978–989.
6. Nakaya,A., Hishigaki,H. and Morishita,S. (1999) Tracing synergetic behavior of the QTLs affecting oral glucose tolerance in the OLETF rat. *Genome Inform.*, **10**, 155–165.
7. Terwilliger,J. and Ott,J. (1994) *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, MD.
8. Abecasis,G.R. and Cookson,W.O.C. (2000) GOLD—Graphical Overview of Linkage Disequilibrium. *Bioinformatics*, **16**, 182–183.
9. Wu,R. and Zeng,Z.B. (2001) Joint linkage and linkage disequilibrium mapping in natural populations. *Genetics*, **160**, 899–909.
10. Devlin,B. and Risch,N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311–322.
11. Hill,W.G. and Robertsoon,A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.*, **38**, 226–231.
12. Hill,W.G. and Weir,B.S. (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am. J. Hum. Genet.*, **54**, 705–714.
13. Lewontin,R.C. (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**, 49–67.